PERSONNEL ASSESSMENT AND DECISIONS

Personnel Assessment and Decisions

Volume 4 | Issue 1 Article 2

2018

Critical Analytic Thinking Skills: Do They Predict Job-Related Task Performance Above and Beyond General Intelligence?

Sara Beth Elson
MITRE Corporation, selson@mitre.org

Robert Hartman
MITRE Corporation, rhartman@mitre.org

Adam Beatty
Human Resources Research Organization, abeatty@humrro.org

Matthew Trippe
Human Resources Research Organization, mtrippe@humrro.org

Kerry Buckley
MITRE Corporation, kbuckley@mitre.org

Critical Analytic Thinking Skills: Do They Predict Job-Related Task Performance Above and Beyond General Intelligence?

Authors

Sara Beth Elson, Robert Hartman, Adam Beatty, Matthew Trippe, Kerry Buckley, John Bornmann, Elaine Bochniewicz, Mark Lehner, Liliya Korenovska, Jessica Lee, Les Servi, Alison Dingwall, Paul E. Lehner, Maurita Soltis, Mark Brown, Brandon Beltz, and Amber Sprenger

Critical Analytic Thinking Skills: Do They Predict Job-Related Task Performance Above and Beyond General Intelligence?

Sara Beth Elson[1], Robert Hartman[1], Adam Beatty[2], Matthew Trippe[2], Kerry Buckley[1], John Bornmann[1], Elaine Bochniewicz[1], Mark Lehner[1], Liliya Korenovska[1], Jessica Lee[3], Les Servi[1], Alison Dingwall[1], Paul E. Lehner[1], Maurita Soltis[1], Mark Brown[1], Brandon Beltz[1], and Amber Sprenger[1]

1. MITRE Corporation
2. Human Resources Research Organization
3. State Department

KEYWORDS

critical thinking, job performance, criterion-related validity, test development

ABSTRACT

Employers and government leaders have called attention to the need for critical thinking skills in the workforce, whereas business trends toward evidence-based decision making also highlight the increasing importance of the critical thinking skill set. Although studies have examined the relationship of critical thinking to behaviors or job performance, many have missed a key component: incremental predictive validity of critical thinking beyond cognitive ability. The current study defines critical thinking, presents results from a test development effort in which the conceptual definition was operationalized as a measure of critical analytical thinking skills for government analysts, and presents results of a criterion validity study examining whether critical thinking skills predict technical performance generally and incrementally, beyond cognitive ability and other characteristics.

In our increasingly knowledge-oriented economy (Powell & Snellman, 2004), employers and government leaders have expressed substantial interest in the notion of "21st century skills," which include critical thinking skills among others (Pellegrino & Hilton, 2015). Business trends toward evidence-based decision making (Buluswar & Reeves, 2014) and the advent of the Big Data movement (Putka & Oswald, 2015) also point to the increasing importance of the critical thinking skill set. For example, Casner-Lotto and Barrington (2006) found that among 400 surveyed employers, 92.1% identified critical thinking/problem-solving as being very important in shaping 4-year college graduates' success in today's workforce, and critical thinking was also considered important for high school and 2-year college graduates. More recently, a survey by the Association of American Colleges and Universities (AAC&U, 2011) found that 81% of employers wanted colleges to place a stronger emphasis on critical thinking. Consistent with this expressed need, several standardized critical thinking tests have been developed (Ennis, Millman, & Tomko, 1985; Ennis & Weir, 1985; Facione, 1990; Facione & Facione, 1992; Halpern, 2010; Paul & Elder, 2006; Watson & Glaser, 2009).

Despite this widespread interest in the cultivation and measurement of critical thinking skills, definitions of the construct are varied (Liu, Frankel, & Roohr, 2014). Markle, Brenneman, Jackson, Burrus, and Robbins (2013) reviewed seven frameworks concerning general education competencies deemed important for higher education or the workforce. They found that, although there is overlap in the frameworks' definitions, there is also variation in what the different frameworks regard as the core features of critical thinking. Similarly, our review of existing critical thinking tests underscored the diverse ways that theorists and test developers have conceptualized critical thinking elements.

Corresponding author:
Amber Sprenger
Email: asprenger@mitre.org
Phone: 703-983-4717

Again, although there was significant overlap across measures, it was frequently the case that a given test instrument would feature one or more subscales that had no direct parallel in the other test instruments.

In addition to this uncertainty surrounding the elements of critical thinking, there is the question of whether critical thinking skills can be distinguished from general mental ability (i.e., GMA – intelligence or general cognitive ability; Hunter & Hunter, 1984; Schmidt & Hunter, 1998) or from general intelligence (i.e., *g*; Jensen, 1998). On the one hand, considerable research supports the "positive manifold" hypothesis that diverse measures of knowledge and reasoning skill tend to be significantly, positively intercorrelated (Hunt, 2011). As noted by Lake and Highhouse (2014), the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 2009), which has a long history of use in organizational hiring and promotions since its development in 1925, diverges in format from conventional intelligence tests but can be expected to relate substantially to measures of intelligence, such as the Raven's Advanced Progressive Matrices (r = .53, Raven & Court, 1998) and the WAIS intelligence test (r = .52, Watson & Glaser, 2009). However, other scholars have argued that general intelligence alone cannot explain critical thinking. For example, Stanovich and West (2008) examined critical thinking skills in eight different experiments. They discovered that participants with high cognitive abilities (as measured by self-reported verbal, mathematical, and total SAT scores) displayed the same level of biases as participants with low cognitive abilities, suggesting that general intelligence does not in and of itself enable people to engage in critical thinking tasks that have been discussed in the literature.

Stanovich, West, and Toplak (2012) have also highlighted dual process models of cognition (e.g., Frederick, 2005) as helping to elucidate the difference between *g*/GMA and critical thinking. Such models posit a distinction between an automatic, heuristic mode of cognitive processing (Type 1) and a slower, more analytic and computationally expensive mode of processing (Type 2). A key distinction between these two processing modes is that, whereas Type 1 processing happens rapidly and relatively automatically, people can make a conscious decision to engage in effortful Type 2 processing, and the willingness to do so can be viewed as a cognitive style. By this conceptualization, g could be considered a form of Type 1 processing, whereas critical thinking could be considered a form of Type 2 processing. On this basis, Stanovich et al. have contended that measures of g (such as IQ tests) do not capture the propensity to engage in effortful, critical thinking.

The question of whether critical thinking is a distinct construct from general intelligence and, in particular, whether it can explain technical performance above and beyond the ability of general intelligence constituted a key impetus for the current study.

Validity of Critical Thinking Measures

Although most studies of critical thinking test validity have focused on correlations with other critical thinking measures or with *g* (Liu et al., 2014), a set of notable studies have examined the relationship of critical thinking to behaviors, job performance, or life events. In their review of literature on the validity of critical thinking measures, Liu et al. (2014) concluded that many existing studies are missing a key component, namely incremental predictive validity of critical thinking above and beyond general cognitive measures. For example, Ejiogu, Yang, Trent, and Rose (2006) found that the Watson-Glaser Critical Thinking Assessment (WGCTA) correlated moderately with job performance (corrected r = .32 to .52). In addition, Watson and Glaser (2009) found that scores on the WGCTA predicted supervisor ratings of judgment and decision-making job performance (r = .23) in a sample of 142 managers across multiple industries. As noted by Lake and Highhouse (2014), judgment and decision-making performance are considered as part of an "analysis" construct, along with "decisiveness" and "adaptivity," which compose three constructs serving as specific, proximal (and ultimately more useful) predictors of managerial decision-making competence than broad constructs like cognitive ability and personality (see Lievens & Chan, 2010). Watson and Glaser (2010) also found that the WGCTA correlated at .40 with supervisor ratings of analysis, problem-solving behaviors, and judgment and decision-making behaviors for analysts from a government agency. Butler (2012) found that scores on a different measure of critical thinking (the Halpern Critical Thinking Assessment or HCTA) predicted real-world outcomes of critical thinking, that is, decision outcomes (as assessed by the Decision Outcomes Inventory (DOI: Bruine de Bruin, Parker, & Fischhoff, 2007). Garrett and Wulf (1978) found that Cornell Critical Thinking Test (CCTT) scores predicted academic success in graduate school, i.e., grade point average (GPA). Finally, Stilwell, Dalessandro, and Reese (2011) found that Law School Admission Test (LSAT) scores predicted GPA for law school students' first year.

Unfortunately, none of these studies assessed whether critical thinking predicted criterion variables above and beyond the ability of general intelligence measures. This represents a significant gap in the critical thinking skills test validity literature (see Liu et al., 2014), because *g* is consistently identified as the single most predictively valid psychometric indicator of individual job performance (Schmidt & Hunter, 1998; see also Heneman & Judge, 2012 on cognitive aptitude). For example, Hunter's (1980) meta-analysis with 32,000 employees in 515 jobs found that *g* and work performance correlated strongly (r = .51), with validity coefficients being highest for higher-complexity occupations (.58 vs. .23 for high vs. low complexity jobs). More recently, Ones, Dilchert, Viswesvaran, and Salgado

(2010) reported operational validities (correlations corrected for range restriction and reliability) between .35 and .55.

Furthermore, studies of incremental predictive validity have underscored the uniqueness and criticality of *g*. That is, previous research has generally found that specific cognitive abilities do not have incremental validity beyond that provided by g (Brown, Le, & Schmidt, 2006; Hunter, 1986; Olea & Ree, 1994; Ree & Earles, 1991; Ree, Earles, & Teachout, 1994; Schmidt & Hunter, 2004; Schmidt, Hunter, & Caplan, 1981; Schmidt, Ones, & Hunter, 1992). Given this lack of research findings, Kuncel (2011) noted that evidence of predictive validity beyond that of *g* will be needed to better assess the unique, marginal benefits of critical thinking tests.

Aims of the Present Research

The current study represents a first step in addressing the conceptual and empirical gaps within the literature. Specifically, we present the outputs of an effort to canvass existing definitions and models of critical thinking skills to arrive at a consensus set of critical thinking elements or subconstructs. In addition, we summarize previously unpublished results from a test development effort, in which our conceptual definition was operationalized as a measure of critical analytical thinking skills for government analysts. Finally, we present the results of a criterion validity study that examined whether critical thinking skills predict technical performance generally and incrementally, above and beyond a measure of *g* as well as above and beyond job experience, educational attainment, and a series of other characteristics.

It should be noted that the current study emerged as part of a broader effort to develop the Critical Analytic Thinking Skills (CATS) test (MITRE Corporation, 2014a; MITRE Corporation, 2015), a measure of critical thinking skills intended for use among government analysts. In particular, the test content was developed specifically to have high face validity for government analysts, which was accomplished by couching the test items in terms of contextualized scenarios. Despite this contextualized framing, items were intended to tap classes of critical thinking skill of broad relevance to any occupation for which such skills are vital. As such, the CATS test can be regarded as an occupation-specific instantiation or translation of a more general purpose conceptual and test item development framework developed over the course of the project. Further, no specialized knowledge of content is required to comprehend the questions and reason to the correct answers.

Elements of Critical Thinking

Given a lack of consensus among researchers on how to define critical thinking and the unique employment context in which we conducted the current study, we pursued several distinct lines of effort to define and operationalize the construct of critical thinking for this context. To identify relevant critical thinking skill elements and refine their definitions, we held a CATS Workshop to elicit perspectives from leading experts in the fields of test development, critical thinking, and analysis (n = 35). In addition, we assessed existing measures of critical thinking and related literature to understand the full scope of the critical thinking construct and various permutations thereof (e.g., Bondy, Koenigseder, Ishee, & Williams, 2001; Ennis & Weir, 1985; Facione, 1990; Frisby, 1992; Halpern, 2010; Klein, Benjamin, Shavelson, & Bolus, 2007; Watson & Glaser, 2010). We gathered additional input from an informal focus group (n = 4) and the CATS Technical Advisory Committee (TAC; n = 8). We also examined critical thinking skill elements included in occupation-specific documents. Finally, we examined 12 government critical thinking training course syllabi to investigate which elements were included as major topics. (Full details of these tasks are discussed in "Critical Analytical Thinking Skills Pilot Test Final Report" [MITRE Corporation, 2014b]). The end products of this effort were a high-level conceptual definition of critical thinking as "the reflective use of cognitive skills to make good judgment" along with an associated set of critical thinking "elements" and element definitions, where an element is a conceptually distinct sub-category of critical thinking skills grouped by similarity.

We initially considered several elements of critical thinking for inclusion in the CATS test. In selecting these elements, we prioritized the need to maximize content validity or the degree to which the test represents all aspects of the critical thinking construct. At the same time, we sought to manage the overall test length. Given these considerations, the final CATS test incorporated four elements with the strongest support from the information sources surveyed: Identifying Assumptions, Causal Reasoning, Logical Reasoning, and Hypothesis Evaluation (see Table 1). Although the primary focus of this report is the assessment of the CATS test's predictive/criterion validity with respect to job performance, a review of prior (previously unpublished) CATS test development and validation work is necessary to help establish the measure's general psychometric properties, including test reliability and convergent validity with other relevant cognitive measures. Therefore, before presenting the core hypotheses for the present effort, we provide a short overview of prior psychometric evidence concerning CATS.

Item Analysis and Scale Construction. A total of 246 multiple-choice items were initially generated by trained item writers to measure the four elements of critical thinking, and 209 survived an expert review process. A pilot study was then conducted to collect item statistics using a sample of Amazon's Mechanical Turk (MT) participants (n = 511). The pilot test sample was restricted to US citizens

TABLE 1.
Elements of Critical Thinking

Element	Definition
Identifying assumptions	Assumptions are statements that are assumed to be true in the absence of proof. Identifying assumptions helps to discover information gaps and to accurately assess the validity of arguments. Assumptions can be directly stated or unstated. Detecting assumptions and directly assessing their appropriateness to the situation helps individuals accurately evaluate the merits of arguments, proposals, policies, or practices.
Causal reasoning	Causal reasoning involves evaluating the likelihood of causal relationships among events or other variables. Good causal reasoning requires understanding the concepts of and differences between causation and correlation. Causal reasoning involves identifying proper comparison groups, understanding the role of randomness for inferring causation, considering the possible presence of confounding variables, and understanding the role of sample size and representativeness for making appropriate causal inferences.
Logical reasoning	Logical reasoning involves identifying logical connections among propositions and avoiding logical fallacies for inductive and deductive inference. These can include fallacious inferences (e.g., conclusions do not follow from premises, reversal of if-then relationships, circular reasoning), fallacies of relevance (e.g., ad hominem arguments), fallacies of ambiguity in language (e.g., equivocation, straw-man fallacy), and fallacies of presumption (e.g., false premises, tautology, false dichotomy). A capacity for logical reasoning protects against belief bias or the tendency to incorrectly evaluate data in syllogistic reasoning because of prior preferences and expectations.
Hypothesis evaluation	Evaluating hypotheses requires the consideration of alternative explanations regarding a range of actual or potential evidence to test their relative strength. Hypothesis evaluation may involve comparing a specific hypothesis against the null hypothesis that nothing special is happening or against one or more competing alternative hypotheses to determine which hypothesis is most consistent with or explanatory of the relevant data.

with at least some college education. The final set of CATS items was selected based on traditional classical test theory statistics such as item difficulty, item discrimination statistics, and interitem correlations. Items deemed eligible for inclusion in the CATS test were diverse in difficulty, highly discriminating, and had good statistics for all distractors, as gauged by the proportion of test takers answering each distractor item correctly (pvals) and by option-total, point-biserial correlations (OTCs) used to identify items for which high ability test takers were drawn to one or more distractors.

To meet the needs of potential test users, three forms of CATS were developed to accommodate practical constraints of testing time: A long form containing 156 items that measured all elements, a two-element test (CATS 2-Short) that consisted of only logical and causal reasoning items, and a four-element short form (CATS 4-Short) that included all four elements. In determining the final test length and composition, key consideration was given to (a) the ability to maximize the test's reliability and content validity, (b) resistance to format effects, (c) ceiling effects, (d) guessing and compromise, suitability for Adaptive Computer Testing, and (e) item response theory (IRT) analyses, and (f) test development costs.

Mean scores, standard deviations, reliabilities, and interelement correlations were calculated for each element and test form. Reliabilities of the test forms were high, ranging from .84 to .96. Element scores were highly correlated with each other and with form scores, suggesting a high degree of homogeneity across elements. Results of a confirmatory factor analysis indicated that the CATS elements were correlated at .9 or higher, indicating that test

interpretation should focus on the overall test score as opposed to using the element subscores, as the results did not support the hypothesis that the elements were unique.

Convergent Validity

After completing the scale construction study, a convergent validity study was conducted to evaluate the test's correspondence with well-established measures of critical thinking, including the Law School Admission Test Logical Reasoning Scale (LSAT LR; Roussos & Norton, 1998) and the Shipley Institute of Living Scale 2 (Shipley 2) Cognitive Ability test (Kaya, Delen, & Bulut, 2012). Based on analysis of data collected using the MT participant sample, the corrected correlations between the CATS elements and the established reasoning tests demonstrated convergent (r = .70 to .90) and discriminant (r = .30 to .40) validity.

Parallel Forms Development

As a follow-up to the pilot study discussed above, we conducted a separate MT study with almost double the number of participants (n = 943) and many newly constructed items. This study had several goals, including (a) confirming the findings of the pilot study, (b) conducting item response theory (IRT) calibration of the CATS items, and (c) developing parallel forms for testing scenarios when equivalent forms are desired.

Results from this follow-up study replicated the findings of the pilot study. The difficulty of CATS 2.0 items ranged widely, the items were reliable, appeared largely to measure one general factor, and had expected patterns of convergent validity with established cognitive ability measures. IRT calibration was successful, with a low percentage of items needing to be dropped due to not fitting the model and exhibiting local dependence.

After completing IRT calibration to obtain the final operational item pool, parallel forms were constructed. A total of three sets of parallel forms, focusing on different ability levels and testing scenarios, were developed. These forms exhibited high internal consistency and test-retest reliability.

Convergent Validity Replication

To determine the convergent validity of the parallel forms, a replication of the Year 1 convergent validity study was conducted, including the LSAT and Shipley-2 test as marker tests. Replicating the Year 1 results, the CATS total and form scores correlated strongly with the LSAT Logical Reasoning subtest (i.e., corrected correlations ranged from .81 to .91, see Table 2), demonstrating convergent validity. On the other hand, discriminant validity evidence comes from the corrected correlations between CATS scores and the *Shipley Block Patterns* test (i.e., .37 - .50), as would be expected given that this test measures a somewhat distinct construct from CATS. Finally, CATS elements and forms were correlated more highly with the LSAT-Logical Reasoning test than with the Shipley Vocabulary or Abstraction tests (for which corrected correlations ranged from .39- 63), thus showing patterns of convergent and discriminant validity.

Although the previous work established the psychometric

TABLE 2.

Correlations Among CATS Scores and Marker Test Scores

Score	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Identifying assumptions	.83	.97	.90	.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.43	.52	.37	.56	.50	.84
2. Causal reasoning	.81	.84	.92	.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.46	.55	.40	.60	.54	.87
3. Logical reasoning	.78	.81	.92	.87	1.00	.97	.96	.98	.99	.97	.99	.39	.63	.50	.63	.59	.81
4. Hypothesis evaluation	.78	.78	.76	.82	.99	.94	.95	.98	.96	.96	.95	.49	.53	.41	.59	.56	.85
5. Total score	.90	.92	.95	.88	.96	1.00	1.00	1.00	1.00	1.00	1.00	.44	.59	.45	.62	.57	.86
6. CATS-A: Form A	.82	.85	.84	.76	.90	.81	1.00	1.00	1.00	1.00	1.00	.43	.56	.41	.59	.53	.88
7. CATS-A: Form B	.83	.85	.83	.77	.90	.82	.81	1.00	1.00	1.00	1.00	.42	.56	.41	.59	.53	.89
8. CATS-S: Form A	.85	.87	.87	.81	.93	.90	.87	.85	1.00	1.00	1.00	.45	.57	.41	.61	.54	.89
9. CATS-S: Form B	.85	.88	.88	.80	.93	.89	.90	.86	.85	1.00	1.00	.44	.60	.45	.63	.57	.88
10. CATS-S Short: Form A	.82	.84	.83	.78	.89	.88	.85	.91	.89	.80	1.00	.43	.60	.43	.62	.55	.91
11. CATS-S Short: Form B	.83	.85	.85	.77	.90	.88	.86	.89	.92	.82	.80	.45	.57	.42	.62	.55	.91
12. Shipley-2: Vocabulary	.35	.37	.32	.38	.38	.34	.33	.37	.35	.34	.35	.76	.28	.13	.79	.68	.47
13. Shipley-2: Abstraction	.39	.41	.49	.39	.47	.41	.41	.43	.45	.44	.42	.20	.66	.61	1.00	.63	.67
14. Shipley-2: Block Patterns	.33	.35	.46	.35	.42	.35	.35	.36	.39	.36	.36	.11	.47	.91	.51	.99	.43
15. Shipley-2: Composite A	.44	.48	.53	.47	.53	.47	.47	.49	.51	.49	.48	.60	.85	.43	.76	.84	.69
16. Shipley-2: Composite B	.42	.45	.53	.46	.52	.44	.44	.46	.48	.45	.46	.55	.48	.87	.67	.85	.57
17. LSAT: Logical Reasoning A	.62	.64	.63	.62	.68	.64	.65	.67	.66	.65	.65	.33	.44	.33	.49	.43	.65

Note. Sample size = 943. Coefficient alpha reliability estimates appear on the diagonal for all variables except Shipley scores. Shipley reliability values are split half reliability estimates, corrected to test length using the Spearman-Brown formula. Correlations below the diagonal are correlations observed in the study. Correlations above the diagonal are corrected for unreliability where $r_{1'2'} = r_{12} / \sqrt{(r_{11} * r_{22})}$. Corrected correlations greater than 1 are reported as 1.00.

soundness of the CATS test, this research was conducted with MT workers, and no relevant criteria were available to determine the criterion-related validity of the test. Therefore, we conducted the present study to examine the extent to which the test might have criterion-related validity – especially when administered to government analysts.

The Present Research: Criterion Validity and Incremental Validity

After establishing the reliability and convergent validity of the CATS test, our next step consisted of determining whether the test – and, ultimately, the construct of critical thinking – predicts job performance above and beyond general intelligence. As such, we conducted a criterion-related validity (CRV) study of the relationship between CATS test scores and a set of performance-related criterion measures. We examined this relationship in a sample of US government analysts. Our research entailed testing three overall hypotheses:

Hypothesis 1: Critical thinking test scores will predict performance on an analytic work sample task.

Hypothesis 2: Critical thinking skills will predict performance beyond the ability of general intelligence to do so.

Hypothesis 3: Critical thinking skills will predict performance beyond a set of individual characteristics, including general intelligence, educational attainment, gender, employment sector (i.e., whether civilian, military, or contractor), job experience related to the analytic work sample task, completion of training in structured analytic techniques, age, motivation on the CATS test, and motivation on the work sample task.

METHOD

Participants

Participants consisted of 140 government analysts from across a range of organizations. A priori power analysis indicated that 125 participants would allow detection of correlations greater than .22 (i.e., at the "small" or greater level; Cohen, 1992) with a power of .8. In addition to participants, 24 supervisory SMEs were recruited from 11 different agencies across the government for purposes of rating analytic products that the participants would provide during the study. All supervisory SMEs had supervisory-level experience and regularly evaluated analytic products of subordinates.

Materials

CATS test. Participants completed the multiple choice CATS test. For this study, half of participants completed Form A, and the other half completed parallel Form B.

Analytic Work Sample Task. In order to provide empirical evidence that scores on the CATS test predict government analyst job performance, an Analytic Work Sample Task (AWST) was developed to closely simulate the work government analysts perform on the job. The AWST materials were developed using a modeling approach with significant input from subject matter experts (SMEs). As part of the task, participants read a short background primer. After reading this background material, participants viewed a dossier of evidence consisting of reports describing simulated events. Then, participants were instructed to write a short report in the style of an analytic work product, which was evaluated by at least three supervisory SMEs using a standardized rubric developed for this project. The supervisory SMEs were all experienced in evaluating products. Their task scores provided a measurement of how well analysts identified assumptions, considered alternative explanations, evaluated the quality of information sources, drew logical conclusions, and reached accurate judgments with appropriate confidence when writing analytic work products. These performance measures are derived from two government publications on the topic of analytic tradecraft and standards for evaluating the quality of analytic products.[1] Further detail on the AWST can be found in Appendix A.

Cognitive ability measure. Our measure of cognitive ability consisted of self-reported Scholastic Aptitude Test (SAT) test scores and self-reported ACT scores. According to Kanazawa (2006), the SAT Reasoning Test (usually known simply as the SAT or the SAT I) is a measure of general intelligence, defined as the ability to reason deductively or inductively, think abstractly, use analogies, synthesize information, and apply knowledge to new domains, akin to Cattell's (1971) fluid intelligence (Gf). Frey and Detterman (2004) found that the total SAT score is an index of cognitive ability because it loads highly on psychometric *g* (see also Unsworth & Engle, 2007). Furthermore, Engle, Tuholski, Laughlin, and Conway (1999) characterized the verbal SAT (VSAT) and quantitative SAT (QSAT) as reflecting a combination of fluid and crystallized abilities. Coyle (2006) correlated scores on the SAT and ACT with performance on three highly g-loaded cognitive measures (college GPA, the Wonderlic Personnel Test, and a word recall task). The *g*, or general, factor is a common element among all tests of mental ability, the first shared factor that is extracted through factor analysis. Coyle performed a factor analysis that showed high g-loading for raw ACT and SAT scores, and the raw scores were significantly predictive of scores on measures of cognitive ability. In a review of existing research, Baade and Schoenberg (2004) looked at 15 studies of academic achievement and IQ. Their review finds a high correlation between a variety of achievement tests (including the ACT) and scores on the WAIS or WISC. Most college bound students take either the Scholastic Aptitude Test (SAT; College Board Tests Inc., 1995) or the American

1 For access to these documents, please contact Amber Sprenger at asprenger@mitre.org

College Test (ACT; American College Testing Program, 1987) as a college entrance requirement. These measures are employed as predictors of future academic success (e.g., American College Testing Program, 1987; College Board Tests Inc., 1995; Wikoff, 1979), and they correlate highly with measures of intelligence (e.g., Wechsler, 1991). One advantage of using ACT and SAT scores rather than an intelligence test is that intelligence tests administered in low-stakes research settings do not reflect true standing on *g*. Rather, in low-stakes settings motivation acts as a third-variable confound that inflates estimates of predictive validity of intelligence for life outcomes (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011). ACT/SAT scores, which are administered in high-stakes settings wherein test results impact college selection decisions, may more accurately reflect intelligence.

In addition, Lohman and Lakin (2011) have suggested that domain-independent reasoning, a hallmark characteristic of Gf, is a key ability that underlies performance on problems that require domain-specific knowledge—that is, Gc. According to Kanazawa (2006), the ACT is a measure of acquired knowledge, akin to Cattell's crystallized intelligence (Gc). For this reason, we incorporated self-reported ACT scores into a composite variable, along with self-reported SAT scores, to operationalize the construct of cognitive ability. For the present study, participants were asked to indicate their ACT score or their total SAT score (math and verbal if they took the version with two subtests used prior to March 2005, or math, critical reading/verbal, and writing if they took the version with three subtests used from March 2005 to present).

Several studies have indicated that the correlation between self-reported SATs and verified SAT scores is in the range of 0.80-0.90 (Cassady, 2001; Kuncel, Crede, & Thomas, 2005), and self-reported scores have been shown to correlate with a third variable to the same extent as verified SAT scores. Stanovich and West (1998) found that the correlation between a vocabulary test and self-reported SAT total scores (.49) was quite similar to the .51 correlation between the vocabulary test and verified total SAT scores in a previous investigation using the same vocabulary measure (West & Stanovich, 1991).

Demographic questionnaire. Participants completed a demographic questionnaire, capturing the following information: Gender, Age, Highest level of education completed, Organizational affiliation, Training received in Structured Analytic Techniques, Employment status (i.e., active duty military, civil service. contractor), Years of service, Rank/grade level at entry and current rank, and Geographic regions worked.

Post-study questionnaire. Finally, participants completed questions indicating how well they felt the CATS test was contextualized for the government, how difficult they found the CATS test and analytic work sample task, how hard they tried on the CATS test and analytic work sample task, and suggestions for improvement.

Procedure

Administration procedure. Materials were distributed either via computer (n = 127) or paper-and-pencil format (n = 13), depending on participating organizations' preference. Test proctors guided participants through each step of the study.[2]

Analytic work sample rating procedure. The principal criterion variables comprised supervisory SME ratings of each participant's one-two page analytic work sample product. To maintain consistency across supervisory SMEs, all supervisory SMEs attended a training session lasting approximately 2 hours. See Appendix A for details on the training sessions. Supervisory SMEs had no access to analysts' CATS test scores so that bias could not affect analytic work sample ratings. Multiple supervisory SMEs rated each product on several discrete dimensions that are central to the task of analysis (i.e., key judgments, referencing, analysis of alternatives, assumptions and judgments, and logical argumentation) using an evaluation rubric (included in Appendix B, "Evaluation Rubric"). In addition to rating work products along these five dimensions, SMEs also provided an overall rating of each product from "Unacceptable" to "Excellent" (i.e., item 6 of the rubric in Appendix B).

To assign supervisory SMEs to work products, we used partial counterbalancing. Each supervisory SME rated 20 analytic work sample products, and each product was evaluated by 2-4 different supervisory SMEs (four analytic work sample products were each rated by two supervisory SMEs; 65 products were each rated by three supervisory SMEs, and 69 products were each rated by four supervisory SMEs). As such, the present study used an ill-structured measurement design (ISMD) wherein supervisory SMEs and participants were neither fully crossed nor nested (Putka, Le, McCloy, & Diaz, 2008). Although at least two supervisory SMEs judged each analytic work sample product, and most products were rated by three of four supervisory SMEs, not all supervisory SMEs scored all participants (i.e., our design was not fully crossed), and neither was there a separate group of supervisory SMEs scoring each participant (i.e., our design was not fully nested). Therefore, to calculate interrater reliability (IRR), we used the *G(q,k)* statistic proposed by Putka et al. (2008) as our primary measure. This statistic resolves problems with traditional estimators, such as Pearson r and the intraclass correlation (ICC), and serves equally well for crossed, nested, and ill-structured designs.

2 Except for seven (7) participants who completed the task in an unproctored setting.

RESULTS

Participant Characteristics

A total of 140 government analysts were recruited and tested for the CRV study. Participants were predominantly male, and had at least a bachelor's degree, with the largest percent having a master's degree or equivalent. The largest percentage of participants were civil service employees. Their average age was nearly 37, and their average SAT and ACT scores were above the average of the general population. Appendix C has tables providing specific participant characteristics.

CATS Test Scores

Out of a possible total score of 32, participants' mean score was 15.5, with a standard deviation of 5.8 and a range from 5 to 29. Scores exhibited a ceiling of 2.8 *SD*s above the mean and Cronbach's α of 0.96.

Criterion-Related Validity Results

Scoring the Analytic Work Sample Task. Supervisory SMEs (n = 24) rated analytic work sample products using the evaluation rubric included in Appendix B: "Evaluation Rubric." Specifically, SMEs rated the products on the following five analytic performance dimensions, each of which contained at least two subcomponent ratings: (1) assumptions and judgments (two ratings), (2) analysis of alternatives (two ratings), (3) logical argumentation (four ratings), (4) key judgments (two ratings), and (5) referencing (two ratings). Appendix A contains a full description of how we derived composite scores. Ultimately, we summed the ratings across all five dimensions. To ensure that each dimension contributed equally to the overall score, we unit weighted each of the dimensions. For example, ratings for dimensions comprising two items were each multiplied by .5, and ratings for dimensions comprising four items were each multiplied by .25. After summing across all weighted items, we calculated final scores by averaging across SMEs to produce a single composite score for each participant. We will call this score the "product dimension rating."

As noted above, supervisory SMEs also provided an overall rating of each product from "unacceptable" to "excellent" (i.e., item 6 of the rubric in Appendix B). To derive a score for each product, we took an average of supervisory SMEs' ratings. We will call this score the "overall product rating." For purposes of testing the hypotheses listed above, we will focus primarily on the criterion variables of product dimension ratings and overall product ratings.

Assessing interrater reliability.[3] We examined interrater reliability with respect to product dimension ratings and overall product ratings. The interrater reliability (IRR) of supervisory SMEs' analytic work sample ratings was good (product dimension ratings: $G(q,k)$ = .77; overall product ratings: $G(q,k)$ = .70).[4, 5]

Quantifying predictive validity. As discussed above, we examined the ability of CATS scores to predict two criterion variables: product dimension ratings and overall product ratings. We took several approaches to examining predictive validity; these included running Pearson correlations (which is how predictive validity has typically been assessed) and Kendall's Tau coefficients, and running a series of hierarchical regressions to allow for controlling the effects of general intelligence. As discussed above, our measure of cognitive ability consisted of self-reported Scholastic Aptitude Test (SAT) test scores and self-reported ACT scores. (See Appendix D for details on how we created the SAT-ACT variable.)

In support of Hypothesis 1, CATS test scores correlated strongly with analytic work sample performance (product dimension ratings: r = .55, p < .01; Pearson r corrected for measurement error = .64; Kendall's Tau = .40, p < .01. Overall product ratings: r = .56, p < .01; Pearson r corrected for measurement error = .68; Kendall's Tau = .41, p < .01; see Table 3).

To test Hypotheses 2 and 3, we ran a set of hierarchical regressions examining the ability of CATS test scores to predict analytic work sample performance above and beyond a set of individual characteristics to do so. Specifically, in these models, we examined the ability of CATS scores to predict product dimension ratings and overall product ratings. In all cases, we found that CATS test scores significantly predicted unique variance in ratings above and beyond all other characteristics examined. One of the most important individual characteristics examined consisted of a combined SAT-ACT variable. CATS scores correlated significantly with the SAT-ACT combined measure (r = .56, p < .001).

Our first model, presented in Table 4, entailed predicting overall product ratings by first entering the combined SAT-ACT variable and then entering CATS test scores. The combined SAT-ACT variable alone (in Step 1) accounted for 10% of the variance in overall product ratings, but a model that included CATS test scores as well as the combined SAT-ACT variable (in Step 2) accounted for an additional 18% of the

3 In no cases did a supervisory SME rate a work sample written by anyone reporting directly to her/him.

4 As recommended by Putka et al. (2008), we estimated the three variance components underlying the calculation of $G(q,k)$ for both the overall ratings and for the composite scores. Regarding the calculation of $G(q,k)$ for the overall ratings, the ratee main effect variance ($\hat{\sigma}^2_T$) was .52, the rater main effect variance ($\hat{\sigma}^2_R$) was .35, and the combination of Ratee x Rater interaction and residual error variance ($\hat{\sigma}^2_{TR,e}$) was .47. Regarding the calculation of $G(q,k)$ for the composite scores, the ratee main effect variance ($\hat{\sigma}^2_T$) was 3.09, the rater main effect variance ($\hat{\sigma}^2_R$) was 1.57, and the combination of Ratee x Rater interaction and residual error variance ($\hat{\sigma}^2_{TR,e}$) was 1.69. As discussed by Putka et al. (2008), partitioning the variance underlying $G(q,k)$ into these sub-components can help establish a meta-analytic database of variance component estimates that are specific to the types of ratings used by organizational researchers and practitioners. Such a database could then be used to support the calculation of $G(q,k)$ in primary studies that preclude its estimation on locally available data, as explained by Putka et al. (2008).

5 At present, SAS syntax is available for calculating $G(q,k)$ and the variance components underlying it (see Putka et al, 2008).

6 Even after excluding the least motivated participants, CATS test scores continued to predict variance in overall supervisory SME scores above and beyond that predicted by the combined SAT-ACT variable. This was true of all regression results conducted.

TABLE 3.

Correlation Matrix

		1	2	3	4	5	6	7	8
1. Overall product rating	Pearson correlation	1							
	N	138							
2. Product dimension rating	Pearson correlation	.899**	1						
	N	138	138						
3. SAT/ACT scores	Pearson correlation	.309**	.373**	1					
	N	87	87	89					
4. Composite CATS scores	Pearson correlation	.555**	.554**	.559**	1				
	N	138	138	89	140				
5. Education	Pearson correlation	.382**	.457**	.261*	.417**	1			
	N	134	134	89	136	136			
6. CATS motivation	Pearson correlation	.070	.096	.008	.197*	.048	1		
	N	134	134	89	136	136	136		
7. AWST motivation	Pearson correlation	.239**	.313**	.065	.190*	.325**	.430**	1	
	N	133	133	88	135	135	135	135	
8. Age	Pearson correlation	.058	.142	.074	.190*	.583**	.073	.140	1
	N	130	130	88	132	132	132	131	132
9. Employment sector	Cramer's V	.449	.823	.859	0.501*	0.48**	.155	.153	0.684**
	N	134	134	89	136	136	136	135	132
10. Focus on AWST topic	Cramer's V	.421	.857	0.853	0.39	0.225	0.182	.269*	0.481
	N	138	138	89	140	136	136	135	132
11. SAT training	Cramer's V	.527	0.832	0.716	0.463	0.259	0.148	0.2	0.607
	N	138	138	89	140	136	136	135	132
12. Gender	Cramer's V	.483	0.781	0.884	0.377	0.188	0.151	0.126	0.53
	N	134	134	89	136	136	136	135	132

Note. ** Correlation is significant at the 0.01 level (2-tailed).* Correlation is significant at the 0.05 level (2-tailed). Employment sector refers to government, military, or contractor. CATS motivation was assessed at the end of the testing session via a question, "How hard did you try on the critical thinking test (i.e., the test with the multiple choice questions)?" AWST motivation was assessed at the end of the testing session via a question, "How hard did you try on the work sample task (i.e., the task that had simulated materials and you wrote an analytic essay)?" Focus on AWST topic refers to whether the participant focus on the AWST topic in their daily work (i.e., Middle East/Asia) vs. other topics. SAT Training refers to whether or not participants had received training in structured analytic techniques. Associations between categorical variables 9-12 are not meaningful in this context but are available on request.

variance.[6]

A look at the standardized beta weights also shows that CATS test scores significantly predicted overall product ratings above and beyond the ability of SAT or ACT scores.

Our second model, presented in Table 5, entailed predicting product dimension ratings by first entering the combined SAT-ACT variable and then entering CATS test scores. The combined SAT-ACT variable alone (in Step 1) accounted for 14% of the variance in product dimension ratings, but a model that included CATS test scores as well as the combined SAT-ACT variable (in Step 2) accounted for an additional 11% of the variance.

A look at the standardized beta weights also shows that CATS test scores significantly predicted product dimension ratings above and beyond the ability of the combined SAT-ACT variable.

In the final set of regression models, we sought to control for a broader set of characteristics – in addition to the SAT-ACT variable - that might predict performance. We provided the full list of characteristics in Appendix C (Participant Characteristics). Table 6 presents the model in which we predicted overall product ratings by entering the variables described above in a first step and entering CATS test scores in the second step. The combination of variables entered in Step 1 accounted for 23% of the variance in overall product ratings, but a model that includes these variables as well as CATS scores (in Step 2) accounted for an additional 13% of the variance.

A look at the standardized beta weights shows that CATS test scores significantly predicted overall product ratings above and beyond the combination of demographic factors discussed above. In fact, CATS scores constituted the only variable that significantly predicted overall product ratings within the entire model.[7]

Our final model, presented in Table 7, entailed predicting product dimension ratings by first entering the same demographic characteristics as above and then entering

7 Note that the variables included in step 1 jointly explained 23% of the variance, and the lack of statistical significance for any one of these predictors could be due to some multicollinearity. The change in the size and direction of the regression coefficient for the SAT-ACT variable suggests there could be some negative suppression in this analysis.

TABLE 4.

Predicting Overall Product Ratings by First Entering SAT/ACT Scores, Followed by CATS Scores

	B	*SE B*	β
Step 1			
Constant	1.93	0.08	
Combined SAT-ACT variable	0.25	0.08	.31**
Step 2			
Constant	0.62	0.30	
Combined SAT-ACT variable	0.02	0.09	.03
CATS scores	0.08	0.02	.51***

Note: $R^2 = .10$ for Step 1, $\Delta R^2 = .18$ for Step 2 ($p < .001$).
* $p < .05$. ** $p < .01$. *** $p < .001$.

TABLE 5.

Predicting Product Dimension Ratings by First Entering SAT/ACT Scores, Followed by CATS Test Scores

	B	*SE B*	β
Step 1			
Constant	0.003	0.1	
Combined SAT-ACT variable	0.65	0.17	.37***
Step 2			
Constant	-2.19	0.66	
Combined SAT-ACT variable	0.27	0.20	.16
CATS scores	0.13	0.04	.39**

Note: $R^2 = .14$ for Step 1, $\Delta R^2 = .11$ for Step 2 ($p < .01$).
* $p < .05$. ** $p < .01$. *** $p < .001$.

TABLE 6.

Predicting Overall Product Ratings by First Entering Demographics, Followed by CATS Test Scores

	B	*SE B*	β
Step 1			
Constant	1.20	0.70	
Combined SAT-ACT variable	0.17	0.09	0.22*
Education level	0.09	0.08	0.16
Active-duty military versus government employee	-0.40	0.25	-0.22
Contractor versus government employee	-0.24	0.25	-0.11
Focus on AWST topic (Middle East/Asia) versus all others	-0.56	0.23	-0.03
Training versus lack of training in structured analytic techniques	-0.32	0.23	-0.15
Self-reported motivation on the CATS test	0.12	0.13	0.11
Self-reported motivation on the work sample task	0.09	0.13	0.09
Age	-0.01	0.01	-0.14
Gender	-0.10	0.18	-0.06
Step 2			
Constant	-0.02	0.72	
Combined SAT-ACT variable	-0.03	0.10	-0.03
Education level	0.08	0.07	0.15
Active-duty military versus government employee	-0.05	0.25	-0.03
Contractor versus government employee	-0.39	0.23	-0.18
Focus on AWST topic (Middle East/Asia) versus all others	-0.26	0.22	-0.12
Training versus lack of training in structured analytic techniques	-0.23	0.22	-0.11
Self-reported motivation on the CATS test	0.03	0.13	0.02
Self-reported motivation on the work sample task	0.06	0.12	0.06
Age	0.0	0.01	0.0
Gender	-0.01	0.17	0.0
CATS scores	0.07	0.02	0.50***

Note: $R^2 = .23$ for Step 1, $\Delta R^2 = .13$ for Step 2 ($p < .001$). * $p < .05$. ** $p < .01$. *** $p < .001$.

TABLE 7.

Predicting Overall Product Ratings by First Entering Demographics, Followed by CATS Test Scores

	B	*SE B*	β
Step 1			
Constant	-2.21	1.47	
Combined SAT-ACT Variable	0.49	0.18	0.29*
Education Level	0.23	0.16	0.20
Active-duty military vs government employee	-0.45	0.52	-0.12
Contractor vs government employee	0.05	0.51	0.01
Focus on AWST topic (Middle East/Asia) vs all others	0.10	0.48	0.02
Training vs lack of training in structured analytic techniques	-0.89	0.49	-0.19
Self-reported motivation on the CATS test	0.08	0.28	0.03
Self-reported motivation on the work sample task	0.39	0.28	0.19
Age	-0.02	0.02	-0.13
Gender	-0.01	0.38	0.0
Step 2			
Constant	-4.12	1.58	
Combined SAT-ACT Variable	0.17	0.21	0.10
Education Level	0.23	0.16	0.20
Active-duty military vs government employee	0.10	0.54	0.03
Contractor vs government employee	-0.19	0.50	-0.04
Focus on AWST topic (Middle East/Asia) vs all others	-0.22	0.47	-0.05
Training vs lack of training in structured analytic techniques	-0.74	0.47	-0.16
Self-reported motivation on the CATS test	-0.08	0.27	-0.03
Self-reported motivation on the work sample task	0.35	0.27	0.17
Age	-0.01	0.02	-0.04
Gender	0.14	0.37	0.04
CATS Scores	0.12	0.04	0.36**

Note: $R^2 = .28$ for Step 1, $\Delta R^2 = .07$ for Step 2 ($p < .01$). * $p < .05$. ** $p < .01$. *** $p < .001$.

CATS test scores. The combination of demographic characteristics (in Step 1) accounted for 28% of the variance in product dimension ratings, but a model that included CATS test scores as well as the demographic characteristics (in Step 2) accounted for an additional 7% of the variance.

A look at the standardized beta weights shows that CATS test scores significantly predicted product dimension ratings above and beyond the combination of demographic factors discussed above.

DISCUSSION

Adding to a burgeoning set of research findings on the importance of critical thinking skills to job performance, the current study demonstrated the difference that these skills make when performing tasks that government analysts perform. As noted above, CATS test scores correlated strongly with analytic work sample performance (product dimension ratings: $r = .55$, $p < .01$; Pearson r corrected

for measurement error = .64; Kendall's Tau = .40, $p < .01$; overall product ratings: $r = .56$, $p < .01$; Pearson r corrected for measurement error = .68; Kendall's Tau = .41, $p < .01$). As a point of reference, Hunter's (1980) meta-analysis with 32,000 employees in 515 medium-complexity jobs found r =.51 between general mental ability and work performance (corrected for reliability and range restriction on the predictor in incumbent samples relative to applicant populations). The value is higher for jobs with higher complexity (.58) and lower for jobs with lower complexity (down to .23). Although the comparison between the current study and the Hunter meta-analysis is not direct, because the current study uses a work sample task whereas the Hunter meta-analysis is based on supervisor ratings of job performance, the Hunter meta-analysis provides an indication of the size of criterion values that are observed when strong predictors of job performance are assessed.

Going a step further, however, the current study demonstrated the incremental predictive validity of critical thinking skills above and beyond a general intelligence measure (i.e., the combined SAT-ACT variable). In doing so, the current study addressed a gap discussed by both Kuncel (2011) and Liu et al. (2014) in the literature on the validity of critical thinking measures, in that many existing studies have not examined such incremental predictive validity.

In addition to finding that critical thinking predicts task performance above and beyond the ability of general intelligence, the current study entailed controlling for a variety of other individual characteristics that might have accounted for task performance. The fact that critical thinking skills accounted for performance on the work sample task above and beyond the combination of individual characteristics further attests to the importance of these skills to performance.

The findings of this study hold implications for both academic researchers investigating the predictors of job performance and for businesses. For academic studies, the findings suggest that it is worth measuring critical thinking in appropriate contexts. For businesses, the findings substantiate the interest shown in critical thinking skills by managers and government leaders (Pellegrino & Hilton, 2015). In particular, the findings suggest the importance of measuring and testing critical thinking skills when taking an evidence-based decision-making approach toward business management (Buluswar & Reeves, 2014). Although the tests developed in the current study were not designed as screening tools, the results of the study suggest the potential benefits of measuring critical thinking skills in the hiring process as well as before and after analytical training – to gauge the effectiveness of that training.

Strengths, Limitations, and Future Research Directions

The current study has certain methodological strengths, such as the extensive efforts taken to define, operationalize, and ensure the validity of the Critical Analytic Thinking Skills (CATS) test as well as the analytical work sample task used as a proxy for analytical job performance.

However, a limitation warrants discussion. Namely, the study included only one operationalization of *g*, that is, self-reported SAT and ACT scores. Although multiple studies point to the high correspondence between recalled and actual SAT scores (Cassady, 2001; Kuncel et al., 2005), future research can and should include more diverse measures of general intelligence.

In addition, the criterion and predictor variables both assessed maximal performance (what participants "can do") rather than typical performance (what participants "will do" on the job). A recent meta-analysis shows that measures of typical and maximum performance are only moderately related ($r = 0.42$; Beus & Whitman, 2012). One open question is the degree to which typical critical analytical thinking on the job is aligned with maximal performance. Although we do not have empirical data on this, the nature of participants' work has "high stakes" implications that may motivate them to work at their maximum capacity. Nonetheless, an important question left unanswered by the current study is whether CATS would be equally predictive of a different type of criterion measure that could capture typical performance, such as supervisor ratings.

As a third limitation, readers might note the conceptual overlap between certain elements of the CATS test and performance measures of the AWST (i.e., identifying assumptions, considering alternative explanations, and drawing logical conclusions), whereas other performance measures of the AWST are not elements of the CATS test (i.e., evaluating the quality of information sources or reaching accurate judgments with appropriate confidence when writing analytic work products). As noted above, the performance measures of the AWST are derived from published standards for evaluating the analytic integrity of written products, and because elements of critical analytic thinking are central to analytic integrity (and therefore encapsulated among these standards), some conceptual overlap exists between the AWST and the construct of critical analytic thinking, as defined in this article. The purpose of the present project consisted of developing a test that would predict aspects of performance specified by government standards that cannot be predicted by intelligence alone. Notwithstanding the partial conceptual overlap between the CATS test and the AWST, it is worth noting that the CATS is a short, multiple choice test, whereas the AWST takes multiple hours to complete. Furthermore, the SMEs who evaluated the work products were not trained in critical thinking but rather were trained in supervising analysts and evaluating their reports. As such, they were evaluating the work products from the perspective of good work generally (as encapsulated by overall product ratings)—and not simply

by the standards of critical thinking.

One could argue that supervisor ratings would be a more effective criterion variable than the AWST. Ideally, and in the future, supervisor ratings will be examined, but there are drawbacks to these. Supervisor ratings are subject to various forms of unreliability or limited validity. For example, they are known to be subjective, agreement across raters is often low, rating processes are often highly unstandardized, employee-supervisor dyads vary significantly in various ways (e.g., the degree to which the members of the dyad work together closely, duration of the dyad relationship, and degree of supervisor experience in making evaluations), and there are significant variations in evaluation processes across organizations and organizational units. In contrast, some psychometricians have argued that work sample tests have the highest fidelity for measuring criterion performance (Borman, Bryant, & Dorio, 2010).

Finally, we note the issue of range restriction (e.g., the mean ACT score is approximately at the 90th percentile, and the standard deviation is substantially smaller than recent normative data would indicate) such that the correlations between the cognitive ability (i.e., SAT-ACT scores) and the criterion variables as well as the correlation between the SAT-ACT scores and CATS scores may have been attenuated. This attenuation, in turn, would have inflated the estimate of the incremental validity of CATS scores. Ordinarily, we would correct the attenuated correlations for the range restriction if suitable range restriction correction values can be found. Although such values can be found for purposes of correcting SAT and ACT scores relative to the general population, it is highly likely that CATS scores are heavily restricted relative to the general population or even high school test-taking population given reasonably high correlations with other cognitive ability tests (along with arguments about developing CATS-type skills in college). Given these circumstances, it would seem unwise to correct SAT-ACT scores back to the general population but leave CATS scores as they are - just because data are available to do so. Proceeding this way would be erring in the other direction and risks attenuating the CATS-criterion correlations relative to the SAT-ACT score-criterion correlations. In short, the concern about range restriction is a valid one for which data are unavailable to make proper corrections, and so we note the concern as a caveat to our findings.

In conclusion, the current study addresses the notion that measures of general intelligence are sufficient predictors of job performance in contexts not requiring perceptual speed or spatial abilities. Namely, the findings suggest that it may be necessary to measure critical thinking skills as well. We hope that this research will motivate additional studies into the possibility that critical thinking skills are distinct from and play a role beyond that of general intelligence in predicting job performance.

REFERENCES

American College Testing Program. (1987). ACT Assessment Program technical manual. Iowa City, IA: Author.

Association of American Colleges and Universities (AAC&U). (2011). The LEAP vision for learning: Outcomes, practices, impact, and employers' view. Washington, DC: AAC&U.

Baade, L. E., & Schoenberg, M. R. (2004). A proposed method to estimate premorbid intelligence utilizing group achievement measures from school records. Archives of Clinical Neuropsychology, 19, 227–243.

Beus, J. M., & Whitman, D. S. (2012). The relationship between typical and maximum performance: A meta-analytic examination. Human Performance, 25(5), 355–376. http://doi.org/10.1080/08959285.2012.721831

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. Organizational Research Methods, 10(4), 689-709.

Bondy, K. N., Koenigseder, L. A., Ishee, J. H., & Williams, B. G. (2001). Psychometric properties of the California Critical Thinking Tests. Journal of Nursing Measurement, 9, 309-329.

Borman, W. C., Bryant, R. H., & Dorio, J. (2010). The measurement of task performance as criteria in selection research. Handbook of Employee Selection, 439-461.

Brown, K. G., Le, H., & Schmidt, F. L. (2006). Specific aptitude theory revisited: Is there incremental validity for training performance? International Journal of Selection and Assessment, 14(2), 87-100.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. Journal of Personality and Social Psychology, 92(5), 938.

Buluswar, M. & Reeves, M. (2014). How AIG moved toward evidence-based decision making. Harvard Business Review. https://hbr.org/2014/10/how-aig-moved-toward-evidence-based-decision-making

Butler, H. A. (2012). Halpern Critical Thinking Assessment predicts real-world outcomes of critical thinking. Applied Cognitive Psychology, 25(5), 721-729.

Casner-Lotto, J., & Barrington, L. (2006). Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce. New York, NY: The Conference Board, Inc.

Cassady, J. C. (2001). Self-reported GPA and SAT: A methodological note. Practical Assessment, Research & Evaluation, 7(12), 1 – 6.

Cattell, R. B. (1971). Abilities: Their structure, growth, and action. Boston, MA: Houghton Mifflin.

Claudy, J. G. (1972). A comparison of five variable weighting procedures. Educational and Psychological Measurement, 32(2), 311-322.

Cohen, J. (1992). A power primer. Quantitative Methods in Psychology, 112(1), 155-159.

College Board/Educational Testing Service (1995). 1995 College Bound Seniors. New York: College Entrance Examination Board.

College Board. (2017). Equivalence tables. New York, NY: Author. Available at https://research.collegeboard.org/programs/sat/data/equivalence

Coyle, T. R. (2006). Test–retest changes on scholastic aptitude tests are not related to g. Intelligence, 34, 15–27.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. Proceedings of the National Academy of Sciences, 108(19), 7716-7720.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. Organizational behavior and human performance, 13(2), 171-192.

Ejiogu, K. C., Yang, Z., Trent, J., & Rose, M. (2006). Understanding the relationship between critical thinking and job performance. Poster presented at the 21st annual conference of the Society for Industrial-Organization Psychology, Dallas, TX.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. Journal of Experimental Psychology: General, 128(3), 309.

Ennis, R. H., Millman, J., & Tomko, T. N. (1985). Cornell Critical Thinking Essay Test. Pacific Grove, CA: Midwest Publications.

Ennis, R. H., & Weir, E. (1985). The Ennis-Weir Critical Thinking Essay Test. Pacific Grove, CA: Midwest Publications.

Facione, P. A. (1990). California Critical Thinking Skills Test manual. Millbrae, CA: California Academic Press.

Facione, P. A., & Facione, N. (1992). The California Critical Thinking Dispositions Inventory. Millbrae, CA: California Academic Press.

Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 25-42.

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or g? The relationship between the Scholastic Assessment Test and general cognitive ability. Psychological Science, 15(6), 373–378.

Frisby, C. L. (1992). Construct validity and psychometric properties of the Cornell Critical Thinking Test (Level Z): A contrasted groups analysis. Psychological Reports, 71, 291-303.

Garett, K., & Wulf, K. (1978). The relationship of a measure of critical thinking ability to personality variables and to indicators of academic achievement. Educational and Psychological Measurement, 38(4), 1181-1187.

Halpern, D. F. (2010). Halpern Critical Thinking Assessment. Modeling, Austria: Schuhfried (Vienna Test System).

Heneman, H.G. III, & Judge, T.A. (2012). Staffing organizations (7th Edition). New York, NY: McGraw-Hill.

Hunt, E. B., (2011). Human intelligence. Cambridge: Cambridge University Press.

Hunter, J. E. (1980). Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB). Washington, DC: US Department of Labor, Employment Service.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. Journal of Vocational Behavior, 29(3), 340-362.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96(1), 72.

Jensen, A. R. (1998). The g factor: The science of mental ability. Westport, CT: Praeger.

Kanazawa, S. (2006). IQ and the wealth of states. Intelligence (34), 593-600.

Kaya, F., Delen, E., & Bulut, O. (2012). Test review: Shipley-2 manual. Journal of Psychoeducational Assessment, 30(6), 593-597.

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The Collegiate Learning Assessment facts and fantasies. Evaluation Review, 31(5), 415-439.

Kuncel, N. R. (2011). Measurement and meaning of critical thinking. Report presented at the National Research Council's 21st Century Skills Workshop, Irvine, CA.

Kuncel, N. R., Crede, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. Review of Educational Research, 75, 63 – 82.

Lake, C.J., & Highhouse, S. (2014). Assessing decision-making competence in managers. In S. Highhouse, R. Dalal, & E. Salas (Eds.), Judgment and decision making at work. New York: Routledge.

Lievens, F., & Chan, D. (2010). Practical intelligence, emotional intelligence, and social intelligence. In J.L. Farr and N.T. Tippins (Eds.) Handbook of employee selection. New York, NY: Routledge.

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment (ETS RR-14-10), Princeton, NJ: ETS.

Lohman, D. F., & Lakin, J. M. (2011). Intelligence and reasoning. In R. J. Sternberg & S. B. Kaufman, The Cambridge Handbook of Intelligence (pp. 419-441). Cambridge: Cambridge University Press.

Markle, R., Brenneman, M., Jackson, T., Burrus, J., & Robbins, S. (2013). Synthesizing frameworks of higher education student learning outcomes (Research Report No. RR-13-22). Princeton, NJ: Educational Testing Service.

MITRE Corporation (2014a). Critical Analytical Thinking Skills (CATS) Test: Parallel form development. (2009-917826-016). McLean, VA: Author.*

* To obtain this report, please contact Amber Sprenger at asprenger@mitre.org

MITRE Corporation. (2014b). Critical Analytical Thinking Skills Pilot Test final report (2009-917826-016). McLean, VA: Author.*

MITRE Corporation. (2014d). Critical Analytical Thinking Skills Work Sample Task (2009-0917826-16). McLean, VA: Author.*

MITRE Corporation. (2015). Critical Analytical Thinking Skills (CATS) Test Criterion-Related Validity Study final report (2015-14120200002-002) McLean, VA: Author.*

Neubert, J. C., Mainert, J., Kretzschmar, A., & Greiff, S. (2015). The assessment of 21st century skills in industrial and organizational psychology: Complex and collaborative problem solving. Industrial and Organizational Psychology: Perspectives on Science and Practice, 8(2), 238-268.

Norsys Software Corporation. (2008). Netica. Version 4.16, Vancouver, Canada. http://www.norsys.com/

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than g. Journal of Applied Psychology, 79(6), 845.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Salgado, J. F. (2010). Cognitive abilities. In J. L. Farr & N. T. Tippins (Eds.), Handbook of employee selection (pp. 255–275). New York: Routledge..

Paul, R., & Elder, L. (2006). The International Critical Thinking Reading and Writing Test: How to assess close reading and substantive writing. Dillon Beach, CA: The Foundation for Critical Thinking.

Pellegrino, J. W., & Hilton, M. L. (2015). Education for life and work: Developing transferrable knowledge and skills in the 21st century. Washington, DC: National Academies Press.

Powell, W. W., & Snellman, K. (2004). The knowledge economy. Annual Review of Sociology, 199-220.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. Journal of Applied Psychology, 93(5), 959.

Putka, D. J. & Oswald, F. L., (2015). Implications of the big data movement for the advancement of I-O science and practice. In S. Tonidandel, E. King, & J. Cortina. (2015).

Big data at work: The data science revolution and organizational psychology. New York, NY: Routledge.

Raven, J. C., & Court, J. H. (1998). Raven's progressive matrices and vocabulary scales. Oxford, UK: Oxford Psychologists Press.

Ree, M. J. & Earles, J. A. (1991). Predicting training success: Not much more than g. Personnel Psychology, 44(2), 321-332.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. Journal of Applied Psychology, 79(4), 518.

Roussos, L.A., & Norton, L.L. (1998). LSAT item-type validity study. Law School Admission Council Technical Report 98-01. Newtown PA: Law School Admission Council, Inc.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of personnel selection methods in psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin, 124(2), 262.

Schmidt, F. L., & Hunter, J. E. (2004). General mental ability in the world of work: Occupational attainment and job performance. Journal of Personality and Social Psychology, 86(1), 162.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalization results for two job groups in the petroleum industry. Journal of Applied Psychology, 66(3), 261.

Schmidt, F. L., Ones, D. S., & Hunter, J. E. (1992). Personnel selection. Annual Review of Psychology, 43(1), 627-670.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. Journal of Experimental Psychology: General, 127(2), 161.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. Journal of Personality and Social Psychology, 94(4), 672.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2012). Judgment and decision making in adolescence: Separating intelligence from rationality. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, & J. Confrey (Eds.), The adolescent brain: Learning, reasoning, and decision making (pp. 337-378).

Stilwell, L. A., Dalessandro, S. P., & Reese, L. M. (2011). Predictive validity of the LSAT: A National Summary of the 2009 and 2010 LSAT correlation studies. Law School Admission Council: LSAT Technical Report 09-03, October, 2009.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. Psychological Review, 114, 104 – 132.

Watson, G., & Glaser, E.M. (2009). Watson-Glaser II Critical Thinking Appraisal: Technical and user's manual. San Antonio, TX: Pearson.

Watson, G., & Glaser, E. M. (2010). Watson-Glaser II Critical Thinking Appraisal: Technical manual and user's guide. Bloomington, MN: Pearson.

Wechsler, D. (1991). Manual for the Wechsler Intelligence Scale for Children (3rd ed.). San Antonio, TX: The Psychological Corporation.

West, R. F., & Stanovich, K. E. (1991). The incidental acquisition of information from reading. Psychological Science, 2, 325–330.

Wikoff, R. L. (1979). The WISC-R as a predictor of achievement. Psychology of the Schools, 16, 364-366.

RECEIVED 02/09/17 ACCEPTED 11/09/17

Appendix A
Further Detail on the AWST

A Bayesian network (BN; Norsys Software Corporation, 2008) model containing the set of probabilistic and causal relationships among the pieces of simulated evidence formed the basis of all information presented in the work sample materials. In a Bayesian network ("Bayes net"), one node (Bayes net entity) is used for each item (real world entity) to model the interactions within a given problem space. Nodes are connected to one another via links that represent causal relationships, and their interactions are determined by a set of conditional probabilities (e.g., if node A is linked to node B, there will be a set of probabilities that node B will express a certain state given the state of Node A). A Bayes net allows for an understandable representation of complex causal relationships as perceived by domain experts. Once developed, a Bayes net allows the computation of numerous interactions across many variables, such as updating the probability of all variables given any combination of evidence items.

The nodes and causal relationships within the model were informed by a series of interactive sessions with multiple SMEs from a variety of organizations, resulting in a model and corresponding scenario that have complexity and face validity. Following the SME-led development of the model, the specific probabilities and parameters within the model were modified to make it easier to use the Bayes net as a "ground truth" model for generating and evaluating performance on test problems. The resulting CATS Bayes net model, therefore, is not intended to be an exact and accurate domain representation but rather a representation that reflects key complex causal relationships in the domain. Consequently, the Bayes net model can be used to generate realistically complex test problems that resemble real world analysis problems.

We piloted the AWST in a study that included 10 MITRE and 8 government subject matter experts (SMEs) with 5 to 33 years of experience. The methodology used to develop and pilot the analytic work sample task is described in detail in technical report, *Critical Analytical Thinking Skills Work Sample Task* (MITRE, 2014d).

Training Sessions for Supervisory SMEs

During training sessions:

- An overview of the CATS test and the criterion validity study were provided,
- An overview of the analytic work sample task was provided,
- The evaluation rubric was introduced,
- Supervisory SMEs used the evaluation rubric to evaluate a sample analytic work sample product selected from the pilot implementation of the analytic work sample materials.

Supervisory SMEs were provided with the same materials as participants, with the exception of the specific simulated reports, in order to simulate a supervisor's general knowledge of a topic when reviewing analytic products. Although the specific simulated reports were not provided to supervisory SMEs, they did receive descriptions of each piece of evidence (type of report and evidence presented within the simulated reports). In addition, supervisory SMEs were provided with the *Analytic Work Sample Rating Tip Sheet*, which described the analytic work sample BN model in depth and highlighted the most influential indicators, the accuracy of various source-types, the prior year's assessment of the problem set, and how outcome likelihoods changed based on the evidence presented. All documents were reviewed with supervisory SMEs to ensure the SMEs were as familiar as possible with the analytic work sample prior to rating actual participant analytic work sample products.

After providing an overview of the analytic work sample task, supervisory SMEs were provided with a sample analytic product with the following characteristics:

1. The product had a mix of good and bad analysis, allowing supervisory SMEs to discuss strengths and weaknesses on each evaluation rubric dimension,
2. Previous supervisory SMEs in the piloting phase of the analytic work sample construction had specifically identified strengths and weaknesses so that these could be discussed in addition to other items supervisory SMEs identified,
3. The product was in a nonstandard format so that supervisory SMEs would not be primed to expect any given format.

Supervisory SMEs spent approximately 15 minutes reading the sample analytic work sample product and entering their ratings into a sample evaluation rubric sheet. Supervisory SMEs then engaged in a group discussion of each rating. This process allowed supervisory SMEs to raise questions and concerns about the evaluation rubric and other analytic work sample materials, and come to a mutual understanding of each element of the evaluation rubric.

After all supervisory SMEs had completed training sessions, they were sent (via email) 20 analytic work sample products to rate, and allowed 4 weeks to complete the rating process. Of the 25 supervisory SMEs who participated in the training sessions, 24 completed all assigned ratings.

Scoring the Analytic Work Sample Task. Supervisory SMEs (n = 24) rated analytic work sample products using the evaluation rubric. Twelve of the evaluation rubric items evaluate five key analytic performance areas: identifying assumptions, analysis of alternatives, logical

argumentation, key judgments, and appropriate citations. Two of the evaluation rubric items asked the supervisors to provide overall ratings: one of the overall analytic work sample product, and one of the critical thinking skills displayed in the product. Each supervisory SME rated 20 analytic work sample products, and each product was evaluated by 2 to 4 different supervisory SMEs (four analytic work sample products were each rated by two supervisory SMEs; 65 products were each rated by three supervisory SMEs, and 69 products were each rated by four supervisory SMEs). See Appendix F for details on scoring the AWST.

Assessing Interrater Reliability.[8] To assign supervisory SMEs to rate participants, we used partia counterbalancing. We examined interrater reliability with respect to two criterion variables: (1) "product dimension ratings" – derived by taking an average (across supervisory SMEs) of each summed, unit-weighted set of scores that supervisory SMEs assigned each analytic work sample product on each of the five dimensions of analytic performance and (2) "overall product ratings," derived by taking an average of supervisory SMEs overall ratings of each analytic work sample product (i.e., item 6 of the analytic work sample evaluation rubric).

Scoring the AWST. Ratings for each evaluation rubric item were converted to a -1 to +1 scale, where -1 was assigned to the worst response option, +1 was assigned to the best response option, and all other response options were distributed evenly throughout. For instance, for the item, "Identifies indicators that, if detected, could validate or refute judgments," never was coded as -1, sometimes was coded as 0, and almost always was coded as +1. Overall ratings were converted to a 0 to +4 scale, where 0 was assigned to the worst response option, and +4 was assigned to the best response option.

A unit weighting approach was used to calculate the product dimension ratings. Previous research has shown that unit weights perform similarly to, or better than, regression weights, particularly when using smaller samples (Bobko et al., 2007; Einhorn & Hogarth, 1975; Schmidt, 1971; Claudy, 1972). Performance on each dimension was weighted equally, and scores on each dimension were summed to calculate the product dimension rating. Because most evaluation rubric dimensions had two items (i.e., analysis of alternatives; assumptions and judgments; key judgments; referencing), but one had four items (logical argumentation), dimension scores were normalized by the number of items on the dimension so that each dimension contributed equally to the overall composite score. For instance, ratings for dimensions comprising two items were each multiplied by .5, and ratings for dimensions comprising four items were each multiplied by .25. After summing across all weighted items, composite analytic performance scores were calculated by averaging across SMEs to produce a single composite score for each participant.

We attempted to maximize consistency across supervisory SMEs by holding the pre-rating training sessions discussed in Appendix E. Importantly, supervisory SMEs were blind to analysts' performance on the CATS test, so that experimenter bias could not play a role in analytic work sample ratings. In other words, supervisory SMEs could not purposefully rate an analytic work sample higher because they knew someone did well on the CATS test, as they were blind to CATS test scores.

The present study used an ill-structured measurement design (ISMD), wherein supervisory SMEs and participants were neither fully-crossed nor nested (Putka et al., 2008). Although at least two supervisory SMEs judged each analytic work sample product, and most products were rated by three of four supervisory SMEs, not all supervisory SMEs scored all participants (i.e., our design was not fully crossed), and neither was there a separate group of supervisory SMEs scoring each participant (i.e., our design was not fully nested). Therefore, to calculate IRR, we used the G(q,k) statistic proposed by Putka et al. (2008) as our primary measure of interrater reliability. This statistic resolves problems with traditional estimators, such as Pearson r and the intraclass correlation (ICC) and serves equally well for crossed, nested, and ill-structured designs.

8 In no cases did a supervisory SME rate a work sample written by anyone reporting directly to her/him.

Appendix B
Evaluation Rubric

1. Assumptions and **Judgments**

a. Identifies indicators that, if detected could validate or refute **judgments**
- i. Never
- ii. Sometimes
- iii. Almost always

b. Is explicit about **assumptions** important to the analysis
- i. Never or almost never
- ii. Sometimes
- iii. Always or almost always

2. Analysis of Alternatives

a. Presents **analysis of alternatives** where appropriate
- i. Yes
- ii. No

b. Requests additional information that would likely yield evidence to help confirm/disconfirm potential alternatives
- i. Yes
- ii. No

3. Logical Argumentation

a. Analytic **judgments** are supported by references to the text
- i. Never
- ii. Sometimes
- iii. Almost always

b. Language and syntax use
- i. Poor (Is unclear, imprecise and obscures key points)
- ii. Acceptable (Writing is clear and conveys key points)
- iii. Excellent (Makes clear and explicit well-reasoned judgments about trends or underlying dynamics shaping key points)

c. Argumentation:
- i. Completely inconsistent on important points
- ii. Some inconsistencies on important points
- iii. No inconsistencies on important points

d. Causal logic:
- i. Never
- ii. Sometimes
- iii. Almost Always

4. Key Judgments

a. Key judgments:
- i. Most key judgments are questionable or wrong.
- ii. Some key judgments are questionable or wrong.
- iii. All key judgments are correct

b. Confidence in key judgments is:
- i. Excessive given the data
- ii. About right given the data
- iii. Too little given the data

5. Referencing

a. Identifies sources used in analysis
- i. Never
- ii. Sometimes
- iii. Almost always

b. Provides information needed to assess sources used in analysis
- i. Never
- ii. Sometimes
- iii. Almost always

6. Overall rating of this product
- a. Unacceptable
- b. Poor
- c. Fair
- d. Good
- e. Excellent

7. Overall rating of critical thinking skills displayed in this product
- a. Unacceptable
- b. Poor
- c. Fair
- d. Good
- e. Excellent

Appendix C
Participant Characteristics

TABLE 10.
Participant Gender, Education, and Employment Status

		Primary study		Supplemental study	
		N	**%**	**N**	**%**
Gender	Male	88	62.9	103	73.6
	Female	48	34.3	34	24.3
	Not reported	4	3	3	2.1
Education	High school diploma, GED, or equivalent	8	5.7	72	51.4
	Some college	15	10.7	42	30.0
	Associate's degree or other 2-year degree	8	5.7	8	5.7
	Bachelor's degree	34	24.3	14	10.0
	Some graduate school	15	10.7	0	0.0
	Master's degree or equivalent	45	32.1	0	0.0
	Doctorate or professional degree	11	7.9	1	0.7
	Not reported	4	2.9	3	2.1
Employment Status	Active duty military	53	37.9	140	100.0
	Civil service	66	47.1	0	0.0
	Contractor	17	12.1	0	0.0
	Not reported	4	2.9	0	0.0
	Total	**140**		**140**	

TABLE 11.
Participant Age, SAT Scores, ACT Scores, Number of Years of Military and Civilian Service[9]

		Primary study			Supplemental study		
		Mean	**SD**	**N**	**Mean**	**SD**	**N**
Age		36.6	11.2	132	20.5	2.4	137
SAT score	<2005	1230	190	66	1081	556	5
	>2005	1732	434	10	1318	629	53
ACT score		28.5	3.9	33	24.8	4.4	46
# Years active duty military service		5.8	4.3	50	0.4	0.5	121
# Years civil service		10.9	7.9	63	N/A	N/A	0

Note. For reference, SAT scores in 2014 had $M = 1497$, $SD = 322$, and ACT scores in 2009 had $M = 21.1$, $SD = 5.1$. SAT scores in 2004 had $M = 1{,}028$, $SD = 160$ [10]

9 Please note that some participants put SAT and ACT scores that fell outside the ranges for these tests, so these participants were not included when reporting descriptive statistics or running analyses involving SAT and ACT scores. In the case of SAT scores, two participants put scores that fell outside the range, and two did not indicate which version of the test they took (whether before 2005 or starting in 2005). Therefore, these two participants had to be discarded from analyses due to our inability to scale their scores appropriately according to whether they took two subtests or three. Five participants who took the ACT had to be discarded from analysis because they put scores that fell out of range.
10 U.S. Department of Education, National Center for Education Statistics. (2016). Digest of Education Statistics, 2015 (NCES 2016-014), Table 226.10. Available at https://nces.ed.gov/fastfacts/display.asp?id=171

TABLE 12.
Current Civil Service Grade Level

	N	%
GS-1 to GS-3	0	0.0
GS-4 to GS-6	1	0.7
GS-7 to GS-9	0	0.0
GS-10 to GS-12	14	10.0
GS-13 to GS-15	48	34.3
SES	1	0.7
Total	64	45.7

Appendix D
Creation of the Combined SAT-ACT Variable

After obtaining participants' reported SAT and ACT scores, we completed several steps to render these scores comparable and on the same scale. As an initial step, we dropped cases in which participants either reported no SAT or ACT scores, or reported scores that fell outside the acceptable range of each respective test. Some participants reported both an SAT and an ACT score, and in those cases, we examined each pair of scores to look for discrepancies (e.g., an extremely high SAT score and an extremely low ACT score, after z-transforming all scores. We used a set of rules, described in detail below, for determining discrepant scores). Our selection processes resulted in dropping 51 participants from the original 140, resulting in a possible maximum of 89 participants for analysis.

In rendering all SAT and ACT scores comparable, we accounted for the fact that the College Board recentered SAT scores in 1995 and revised the test in 2005 to make the composite scale 600–2400 instead of 400–1600. Our data collection occurred in 2015, before the College Board re-designed the SAT again in the spring of 2016 to revert to the scale of 400-1600. Taking all factors into account, our participants' test scores fell into one of four categories: (1) SAT scores from before 1995, (2) SAT scores from 1995-2004, (3) SAT scores from 2005-2015, and (4) ACT scores. As such, our first step consisted of recentering SAT scores from before 1995 to render them comparable to SAT scores from 1995-2004. Doing so reduced the number of categories from four to three. Our next step consisted of standardizing scores within each of these three subgroups to convert them to z-scores. In the sections that follow, we will detail the specific processes involved in each of these steps.

Converting Scores From Before 1995 to the Recentered Scale. As described by the College Board (2017), "In April 1995, the College Board re-centered the score scales for all tests in the SAT Program to reflect the contemporary test-taking population. Re-centering reestablished the average score for a study group of 1990 seniors at about 500—the midpoint of the 200-to-800 scale—allowing students, schools, and colleges to more easily interpret their scores in relation to those of a similar group of college-bound seniors." Using the College Board's equivalence table, found at https://research.collegeboard.org/programs/sat/data/equivalence/sat-composites we recentered composite SAT scores from before 1995 to place them onto the same scale as scores obtained from 1995 on.[11] To determine which scores pre-dated 1995, we used participants' age as a proxy for test administration date and assumed they had taken the test at age 16. Given that we collected the data during the year 2015, participants who were 36 in that year would have been the first cohort to have their scores re-centered by the College Board. As such, we recentered the scores of participants age 37 and older. After recentering the scores of those participants, our next step consisted of standardizing the scores of our – now – three groups of participants: (1) those who took the SAT before 2005, (2) those who took the SAT between 2005 and 2015, and (3) those who took the ACT.

Standardizing SAT and ACT Scores. Treating each of the three groups listed above separately, we z-transformed all scores – normalizing them only against other scores within each group. In some cases, participants took both the SAT and ACT, and for these participants, we took an average of their z-transformed SAT and ACT scores to derive a single z-score. However, among the participants who took both tests, some got extremely discrepant SAT and ACT scores – after standardization (e.g., an extremely high z-transformed SAT score and an extremely low z-transformed ACT score). It is possible that these participants mistakenly indicated the wrong version of the SAT they took (e.g., if someone indicated they took the SAT before 2005 but reported a score of 2000- when only went to 1600 before 2005). To handle such cases of discrepancy, we applied the following standard: If the z-transformed SAT and ACT scores differed in direction (i.e., positive versus negative) and by more than a standard deviation, we dropped these cases. This procedure resulted in dropping three participants – among the total of 51 dropped (as described above).

In the final set of steps, we combined – into one variable - all the z-transformed SAT and ACT scores as well as the average z-scores for those who had taken both the SAT and ACT. In this manner, we derived our combined, standardized SAT-ACT variable.

11 On its website, the College Board advises researchers that they cannot use the table to convert original V+M scores for a student to recentered V+M scores. Rather, the College Board advises researchers first to convert the student's verbal and math scores from the original to recentered scale using the SAT I Individual Score Equivalents table, and then combine the scores to create a recentered composite. Our protocol did not entail asking participants for their verbal and math scores – only for their composite V+M scores, and so we were compelled to use the table to convert original V+M scores to recentered V+M scores. However, the pattern of correlations (and noncorrelations) between our combined SAT–ACT variable and other variables suggests that our SAT–ACT variable exhibited convergent and divergent validity (see Appendix E for the correlation matrix). Given this finding, we believe our results are valid.